Weekly Report

1 Done

1.1 Vast Presentation

I wrote the script this week. In next weeks, I will keep preparing for presentation.

1.2 Discussion with Zhan Qin

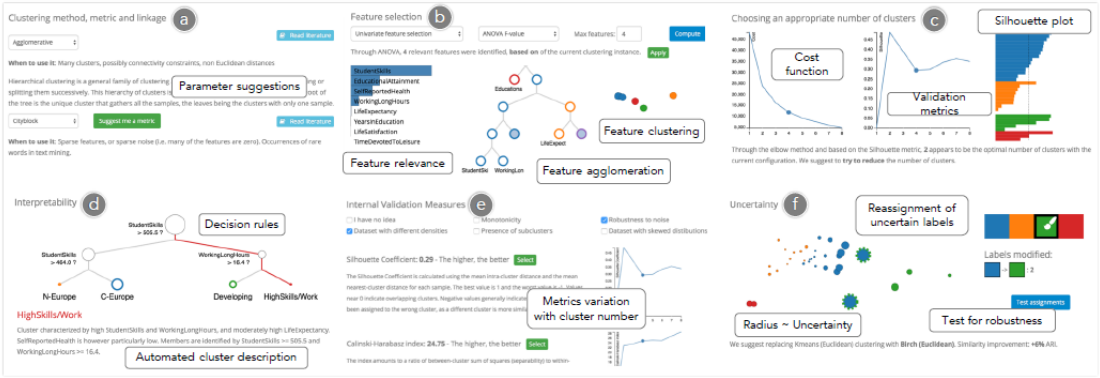
I discuss my idea with him. He replied that he would read related paper first.

1.3 Paper Reading

- *Clustrophile 2: Guided Visual Clustering Analysis*

Clustrophile 2 is a new interactive tool for guided clustering analysis.
Authors summarized nine design criteria based both on research literature and on the regular feedback from data scientists. Among them, some criteria are useful to us:

- Allow quick iteration over parameters.
- Support analysis of large datasets.
- Support reasoning about clusters and clustering instances.
- Guide users in clustering analysis.



Especially the last one. Clustrophile 2 provides guidance for users. The above figure displays a subset of the views included the “Help me decide” (top row) and the “Is this a good clustering?” (bottom row) panels of each Clustering View. Textual explanations and hyperlinks are used to suggest clustering parameters, different feature selection algorithms and visualizations are used to understand the relevance of data dimensions, and cost function and metric plots are used to suggest a good number of clusters. To evaluate the “goodness” of a clustering, decision rules and automated cluster descriptions are used to foster interpretability, several evaluation metrics are dynamically suggested, and uncertain clustering assignments are visualized and tested.

- Structure-aware Fisheye Views for Efficient Large Graph Exploration

This work is from Yunhai Wang's team. They extended the previous research results – structure-based constraints. In the previous research, they employ the constraints to improve the large graph layouts and facilitate in interactive exploration.

This work is about large graph exploration as well. This time, they focus on fisheye effects. It is based on an optimization objective whose terms correspond to the constraints of structure, readability, and temporal coherence.

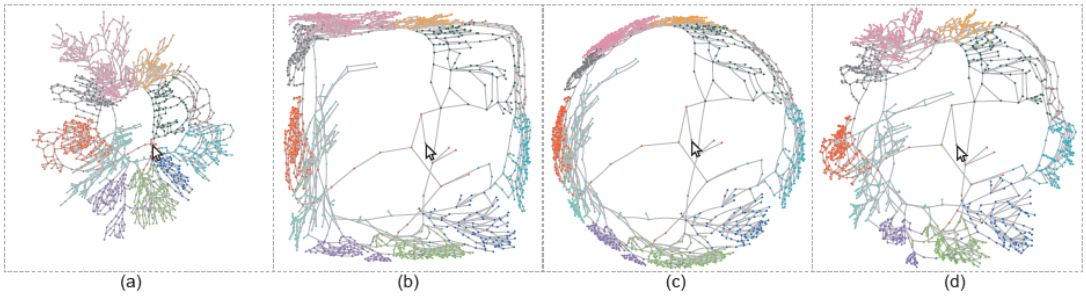
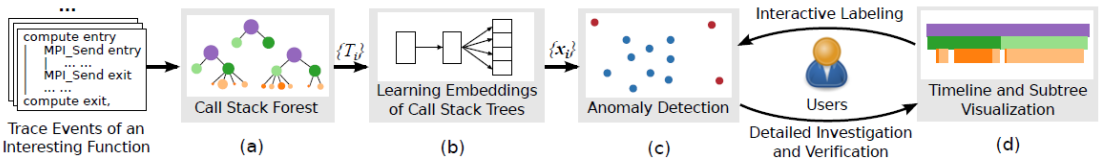


Fig. 1. Magnifying a node-link diagram (a) with 11 clusters around a user-specified location (indicated by the cursor) using different fisheye lenses: (b) graphical fisheye; (c) hyperbolic fisheye; and (d) our structure-aware fisheye, which aims to maintain the shapes of almost all clusters and to minimize their distortions, such as in (b,c).

- A Visual Analytics Framework for the Detection of Anomalous Call Stack Trees in High Performance Computing Applications (VAST HM)

This work is about Anomaly detection.



They first generate CSTrees from the trace events. Next, they construct feature vectors using stack2vec. The candidate anomalous CSTrees are detected in the forest. Then, users can investigate the candidates in detailed visualization. Labeling information provided by the user will be fed back to update the anomaly detection model.

- Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy

Local differential privacy (LDP) techniques collects randomized answers from each user, with guarantees of plausible deniability; meanwhile, the aggregator can still build accurate models and predictors by analyzing large amount of such randomized data.

DEFINITION 1 (ϵ-LOCAL DIFFERENTIAL PRIVACY). *A randomized function f satisfies ϵ-local differential privacy if and only if for any two input tuples $t, t' \in Dom(f)$ and for any possible output t^* of f, we have:*

$$\Pr[f(t) = t^*] \leq e^{\epsilon} \times \Pr[f(t') = t^*].$$

Upon receiving the perturbed data, the aggregator simply computes the average value for each attribute over all users, and outputs these averages as the estimates of the mean values for their corresponding attributes. One of the approaches is calculating an unbiased estimator of the mean. However, the yielded data is biased. The improved method is:

Algorithm 2: Proposed Method for Handling Numeric Attributes

input : tuple $t_i \in [-1, 1]^d$ and privacy parameter ϵ.

output: tuple $t_i^* \in \left\{-\frac{e^\epsilon+1}{e^\epsilon-1}d,\ 0,\ \frac{e^\epsilon+1}{e^\epsilon-1}d\right\}^d$.

Let $t_i^* = \langle 0, 0, \ldots, 0\rangle$;
Sample j uniformly at random from $[d]$;
Sample a Bernoulli variable u such that

$$\Pr[u = 1] = \frac{t_i[A_j] \cdot (e^\epsilon - 1) + e^\epsilon + 1}{2e^\epsilon + 2} \quad ;$$

if $u = 1$ **then**
  $t^*[A_j] = \frac{e^\epsilon+1}{e^\epsilon-1} \cdot d$;
else
  $t^*[A_j] = -\frac{e^\epsilon+1}{e^\epsilon-1} \cdot d$;
return t_i^*

Algorithm 3: Bassily and Smith's method [2]

input : $t_i[A_j] \in [k]$ for each user u_i, privacy budget ϵ, confidence of the error bound β

output: Frequency estimate for each of the k values in attribute A_j

Compute $\gamma = \sqrt{\frac{\log(2k/\beta)}{\epsilon^2 n}}$ and $m = \frac{\log(k+1)\log(2/\beta)}{\gamma^2}$;
Generate random matrix $\Phi \in \{\pm\frac{1}{\sqrt{m}}\}^{m \times k}$;
for $i = 1$ *to* n **do**
  User u_i: draw $s \sim$ Uniform$(\{1, 2 \ldots, m\})$;
  User u_i: draw $t \sim$ Bern$(\frac{e^\epsilon}{e^\epsilon+1})$;
  User u_i: **if** $t = 1$ **then** $\alpha = c_\epsilon m \Phi[s, t_i[A_j]]$ **else** $\alpha = -c_\epsilon m \Phi[s, t_i[A_j]]$, where $c_\epsilon = \frac{e^\epsilon+1}{e^\epsilon-1}$;
  User u_i: submit $\langle s, \alpha\rangle$, which represents a k-dimensional vector z_i where the s-th entry is α and the other entries are 0;
Compute $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$;
for $l = 1$ *to* k **do**
  Estimate the frequency of the l-th value by the inner product of the l-th column of Φ and $\bar{z}$;
return k frequency estimates obtained above;

1.4 File Recovery

With the experience, I uploaded all paper source files to overleaf.

2 Work Hours

In all weekdays, I worked during 9:00~11:30, 13:30~5:00 and 18:30~20:30. On weekends, I worked during 11:30~5:00.

3 Progress

Item	Deadline	Current progress	Remark
Vis presentation	10.24	Script is done.	
Go abroad	11.18	Ready for buying the ticket.	
Privacy program	10.31	Implementing existed approaches.	